Integrity360

your security in mind

Q&A follow-up

AI Risk Management in practice: From Governance to Testing

September 10th, 2025

1. Is there any uptake or adoption of Blockchain technology in AI to address the risk of Data privacy concerns and /or regulatory compliance

   A: I'm not a blockchain expert, but from what I know there are different projects exploring application of blockchain into AI.
   For instance, INATBA (the International Association for Trusted Blockchain Application) has a specific task force for AI and blockchain convergence; take a look at this link: AI & Blockchain Convergences Task Force - INATBA Industrial Blockchain Task Force

2. In your experience, what are the biggest challenges in operationalising AI governance?

   A: The panel identified several major challenges based on their experience:

   Lack of cross-functional collaboration
   AI governance cannot be treated as solely an IT or security issue. Successful implementation requires involvement across the entire organisation, including legal, compliance, HR, and business units.

   Management buy-in
   Without executive sponsorship, AI governance initiatives often fail. Leadership must allocate resources and support the strategic direction of AI adoption.

   Clear ownership and role definition
   Organisations must define who owns AI risk and governance responsibilities. This includes establishing committees or governance bodies to oversee ethical and secure AI use.

   Continuous education for end users
   End users are often the weakest link in cybersecurity and AI governance. Ongoing training is essential to ensure responsible use of AI tools.

   Navigating regional regulations
   For multinational organisations, complying with varying legal frameworks across countries adds complexity. AI governance must be adaptable to different jurisdictions and evolving regulations.

3. What is your opinion on implementing Microsoft CoPilot without MS Purview in place to support Governance?

© Integrity360

A: Yes, it is technically possible to implement Microsoft Copilot without Purview, but it is strongly recommended to use Purview if it's available as part of your licensing. Purview plays a critical role in governance, risk management, and reporting around the use of LLMs, generative AI, and AI tools within a business. One of the key risks Copilot presents is the classification of data created from source data. Purview helps mitigate this by allowing organisations to apply restrictions and controls through data classification. This ensures sensitive information is handled appropriately and supports compliance with internal and external data governance policies. If Purview is not available, organisations should consider alternative tools that offer similar capabilities to manage data security and governance effectively.

4. Can pentesting actually keep pace with how fast models are updated?

A: Yes, but only by shifting to continuous validation. The challenge is that AI patch cycles are rarely visible or documented, as vendors often push silent updates without notice. This means model behaviour can change overnight, and exploits that work one day may stop working the next. Effective testing should not depend on a single lucky jailbreak but on demonstrating broader classes of exploitation. NIST's Gen AI Profile recommends ongoing, lifecycle-based risk management and continuous re-evaluation as models and guardrails evolve, much closer to regression testing than to a one-off audit.

5. Prompt injection is top issue but what is the biggest blind spot when testing AI assistants in 2025?

A: The highest impact comes when LLMs can do things (create tickets, call APIs, touch cloud resources). Research on indirect prompt injection (IPI) shows that poisoned external content can coerce tool-integrated agents into harmful workflows, the blast radius lives after the model output. Public write-ups show models can be steered by hidden or hostile content; this reinforces the need to test retrieval, browsing, and integration paths. AI pentesting should always simulate agentic misuse to avoid aforementioned risks.

© Integrity360

6. How do you measure severity of AI model vulnerabilities?

   A: Start with conventional scoring (e.g. CVSS) but extend with model-aware dimensions (e.g. OWASP LLM Top 10 categories, or NIST's GenAI profile). A jailbreak that just elicits toxic text might be "medium" (content risk). The same jailbreak that results in database queries or financial transactions is "critical" (operational risk). Context of downstream impact is everything.

7. Will the slides be shared after the presentation please?

   A: Yes! They will be included in a follow up email.

8. What do you recommend when it comes to standards / frameworks? How do we select a model?

   A: We can merge these 2 questions. It depends by your objective and business requirement. If you have a business request for a trusted AI system, and you must reassure your stakeholders, the ISO42001 with the related certification is the best solution. NIST AI RMF can be helpful if you are not interested in certification and probably is more flexible than ISO. If you are developing an AI model/system for the EU region, the EU AI Act is mandatory. And let me also add that the EU AI Act came into force in August 2024, but the adoption is gradual, and the full application is planned for August 2027. So, it's quite normal to implement at least two frameworks.

© Integrity360